5 A Survey of Feature Selection in Internet Traffic Characterization*

Bo Zhu and Alberto Mozo

Escuela Técnica Superior de Ingeniería de Sistemas Informáticos

Carretera de Valencia Km. 7 Madrid (28051)

Universidad Politécnica de Madrid

{bozhumatias@ict-ontic.eu, a.mozo@upm.es}

Abstract: In the last decade, the research community has focused on new classification methods that rely on statistical characteristics of Internet traffic, instead of previously popular port-number-based or payload-based methods, which are under even bigger constrictions. Some research works based on statistical characteristics generated large feature sets of Internet traffic; however, nowadays it's impossible to handle hundreds of features in big data scenarios, only leading to unacceptable processing time and misleading classification results due to redundant and correlative data. As a consequence, a feature selection procedure is essential in the process of Internet traffic characterization. In this paper a survey of feature selection methods is presented: feature selection frameworks are introduced, and different categories of methods are briefly explained and compared; several proposals on feature selection in Internet traffic characterization are shown; finally, future application of feature selection to a concrete project is proposed.

Keywords: feature selection, Internet traffic characterization, big data

(*)This research was supported in part by project ONTIC-619633 funded by the European Commission under the Seventh Framework Programme.

5.1 Introduction

In recent years, the world has witnessed an explosion of available information in almost every domain. The sharp increase of the scale of data sets poses a great challenge for scientific researchers when they try to characterize data under research and extract useful knowledge at an acceptable cost. Features are used to convey information as measurable properties to characterize certain aspects of objects under observation in data analysis, machine learning etc. Due to the fast progress of hardware and storage technologies, the scale of feature sets has raised from tens to thousands or even more. In the case of Internet traffic characterization, besides old sample features like protocol category, complex features like Fourier transform of the inter-arrival time of packets [1] are also considered in the latest research works. Handling such a big feature set could be computationally expensive; furthermore, irrelevant and redundant features may also decrease the accuracy of characterization results; finally too many features can severely jeopardize the interpretability of results. As a consequence, feature selection serves as a fundamental procedure of preprocessing in big data scenarios before stepping forward to further application of statistical or machine learning techniques. The main objective of feature selection is to select a subset of features as simplified as possible without suffering a significant decline of accuracy for classification or forecasting, i.e. experimenting with a subset in place of full feature set results in equal or better classification accuracy. The process of feature selection can be totally supervised. In many existing research initiatives domain experts were required to provide a candidate subset of features considering possible domain relevance. However, nowadays due to the large size of feature sets, purely manual feature selection becomes infeasible. Consequently, various feature selection methods have been proposed to generate core feature subsets utilizing different theories and techniques. Feature selection can provide several advantages: It significantly reduces the computational burden, which in turn increases the performance (running time, precision etc.) of classification or prediction models; By getting rid of redundant, irrelevant features or even noise, further processing results don't suffer from bad impacts brought by these interfering

This research was supported in part by project ONTIC-619633 funded by the European Commission under the Seventh Framework Programme.

factors, hence their simplicity and interpretability is improved; It avoids the overfitting problem by building models with only core features, which increases the quality of classification models [2]; As a preprocessing step, feature selection can also help to deeply understand target data sets. Together with domain knowledge, the potential candidate features can be double-validated. The rest of the article is organized as follows: in section 2 we provide an introduction to feature selection frameworks and different categories of feature selection methods as well as a thorough comparison between them; section 3 briefly introduces several feature selection research works in the domain of Internet traffic characterization. Finally in section 4 we propose a future application of feature selection to the ONTIC project and draw conclusions.

5.2 Feature Selection

5.2.1 1.1 Feature Selection Framework

In [3],[4] a basic framework of feature selection was proposed by H. Liu et.al, which is shown in 5.1.

From the figure above, we can see that the feature selection process contains four key procedures: feature subset generation, candidate subset evaluation, stopping criteria judgment and result validation. Previous feature selection research proposals distinguish themselves from each other mainly based on the first two procedures. Since different searching strategies are applied during the generation process, various candidate feature subsets will be available utilizing different methods, and therefore the subset quality and running time performance also vary. Similarly, the optimal feature subset under one evaluation measurement may not be the best option when considering another measurement. This will be discussed in detail in the following section. As mentioned in [5], some previous work focused on the removal of irrelevant feature, but failed to handle redundancy problems. Another framework, shown in 5.2, was proposed in [6] to address the negative effect of redundancy on the speed and accuracy of learning algorithms. First they defined a synthetic function f based on a series of metrics such as irrelevance, redundancy, relevance, sample size, etc. A smaller data set was

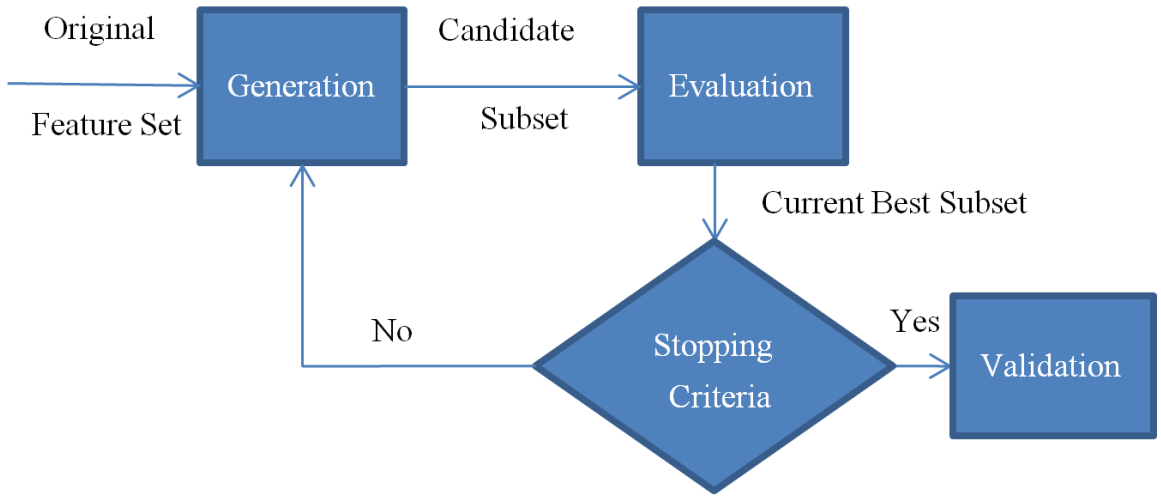

Figure 5.1: Basic framework of feature selection [3]

obtained taking into consideration function f, on which different feature selection algorithms were conducted, resulting in a hypothesis H. Finally a score was generated comparing the defined function and the obtained hypothesis using several predefined scoring criteria.

Since searching through possible candidate feature subsets in the generation process is computationally costly, in [7] H. Liu et.al proposed an improved framework of feature selection to avoid initial subset search step and explicitly eliminate redundant features. This framework is shown in 5.3. In [5], a similar framework was also proposed, which first removes irrelevant features using T-Relevance as a relevance measure quantified by symmetric uncertainty, and then eliminates redundant features by means of a novel feature selection algorithm named FAST.

5.2.2 Feature Selection Method Categories

Previous works on feature selection categories distinguish themselves mainly based on the different search strategies used in the subset generation process [8] or evaluation measure-

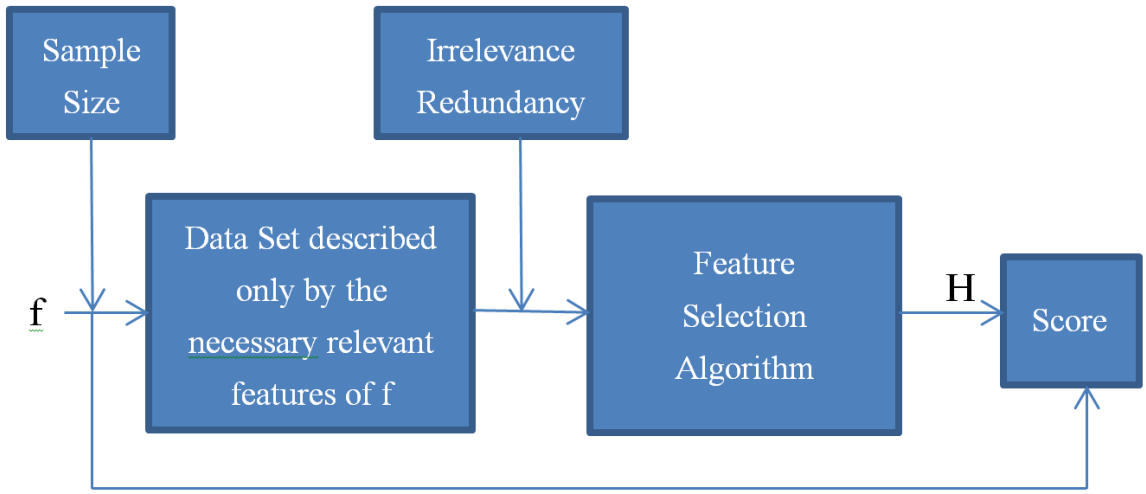

Figure 5.2: Framework of feature selection based on irrelevance and redundancy [6]

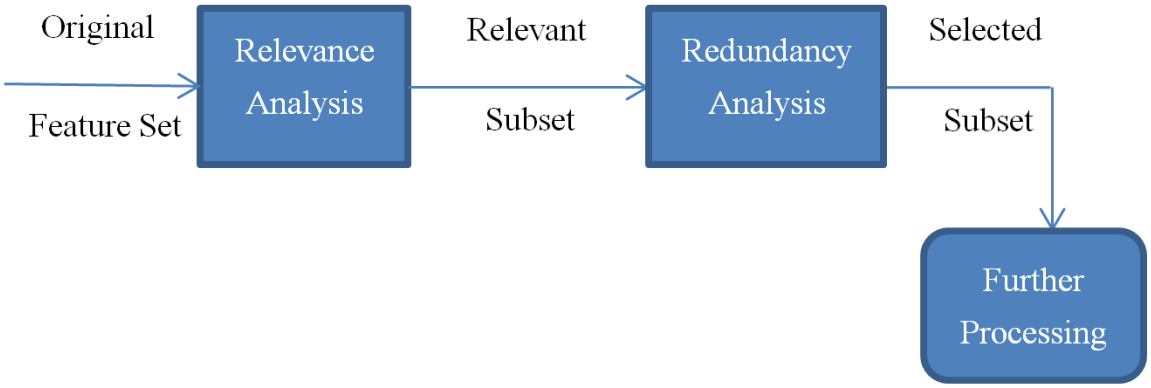

Figure 5.3: Improved framework of feature selection [7]

ments considered in the subset evaluation process [9], or both of them at the same time [3].

Search Strategy In terms of search strategy, feature selection methods can be generally cat-

egorized into 3 types:

Complete Search: This kind of search strategy carries out a complete traversal within the whole feature space, which can be exhaustive or non-exhaustive.

Exhaustive search is simple and intuitive, but suffers from a heavy computational cost. Search methods like breadth first search belong to this type. The time complexity is $O(2^n)$, which makes it infeasible to apply to feature sets of large scale.

Non-exhaustive complete search includes a heuristic process with a backtracking step; once certain judgment criteria are satisfied, the backtracking will be triggered. This ensures that the global optimal subset will be obtained while reducing search space. Examples are:

- Brand and Bound Search: Subproblems are further divided into subproblems until feasible solutions better than the current solution are found; otherwise, this subproblem is pruned.

- Beam Search: An improved variant of B & B Search [10]. At first, the K best features are estimated using scoring criteria. Downstream search starts from these K features by adding one more feature, and K feature subsets with highest score are stored in the queue. In this way, only the most promising K features are considered at each depth level, which significantly reduces search space.

- Best First Search: Similar to Beam search, at each depth level only the feature with the lowest value of the evaluation function is selected for further expansion.

Heuristic Search:
Heuristic Search carries out the search process within the state space. It assesses every candidate and obtains the best one. Then the search process starts over from this temporary point

until the optimal result is finally achieved. Thus many search paths can be discarded directly, which improves its performance greatly. The goodness of each candidate is evaluated by an assessment function established with the help of domain-specific information or heuristic information.

- Sequential Forward Selection: Forward methods start with empty feature subset F. Each iteration, the feature f that makes the value of feature objective function F(f) optimal is added to F. This process is repeated until the number of features of the expected subset is reached or the predefined classification accuracy threshold is achieved. The main drawback of naÃ¯ve sequential selection methods is the inference of nested subsets. Since the correlations between features are not taken into consideration, if feature A is entirely dependent on feature B and both features are added into the same subset, then a nested subset with redundancy is generated.

- Sequential Backward Selection: As opposed to SFS, backward selection starts from the full feature set, and each iteration it discards the feature that makes the feature objective function optimal after elimination. The process goes on until stopping criteria are satisfied. Plus-L Minus-R Selection: This is an improvement based on naive sequential selection. Starting from the empty set, each iteration L features are added, then R (L ¿ R) features are eliminated to make the feature objective function optimal. The backward LRS method also exists. LRS methods provide a way to consider correlation between features, and consequently avoid nested subsets. The choice of parameters L and R has a crucial impact on algorithm performance and should be selected with caution.

- Sequential Floating Forward Selection: SFFS improves forward LRS, but in contrast, the L and R in SFFS are not constant, i.e. SFFS starts from the empty set, and each iteration it chooses the subset f from unselected features that makes the value of the feature objective function optimal after adding f. Then from chosen subset f, it elim-

inates a subset of f (namely g) that makes the value of the feature objective function optimal after eliminating g.

Random Search:
Random search was introduced into feature selection because the scale of feature sets keeps increasing so sharply that the computational cost using previous search strategies becomes unacceptable. Random search is a simple improvement of greedy search that obtains an optimal feature subset by random sampling. This can greatly reduce computational complexity, which consequently increases efficiency and precision. However, one defect of random search is that most methods get trapped easily in local optimal solutions, but multi-repetition could help to overcome this problem. The initialization of pre-established parameters with proper values is another crucial task for random search methods.

- Simulated Annealing: As an improvement of classic hill-climbing search, it introduces a probability function which can give a probability value to determine whether to choose a generated subset as current best subset. This provides the advantage to jump out of local optimal subset at certain probability.

- Genetic Algorithm: First several feature subsets are generated randomly, for which fitness values are calculated [11]. Then operations of crossover and mutation are conducted with a certain probability, producing the next generation of subsets. The reproduction process follows rules of nature, i.e. subsets with higher fitness are more likely to be selected to generate offspring subsets. The optimal feature subset could be obtained after N (sufficiently large) generations.

- Particle Swarm Optimization: PSO method is very similar to GA method, but it replaces crossover and mutation operators with the control of the accelerating speed of particles (feature subsets in this case) [12]. Two optimal values are tracked by each particle: pbest (optimal value found by the particle itself) and lbest (current optimal value found by any particle among the neighbors of the particle).

Some concrete search methods can adopt two search strategies at the same time. For example, the ant colony optimization algorithm is a random search method as well as a heuristic search method.

Evaluation Measurement Evaluation measurement is used to evaluate the goodness of feature subsets provided in a subset generation process. In terms of operating functions, feature selection methods can also be categorized into 3 types:

Filter Method:
Filter methods judge the goodness of subsets via analysis of internal characteristics of the selected features. Various ranking criteria assess different characteristics among features by generating a score for each feature. A corresponding ranking list illustrates the fitness of every feature regarding certain criteria. A threshold can be pre-established to filter out features with low scores. This type of feature selection methods is independent from training process and specific inductive algorithms [13]. As a result, filter methods can be promoted among different algorithms with low computational requirements. Thus filter methods can be utilized to pre-reduce plenty of obviously irrelevant features, but useful features might also be filtered. It can be further classified into 4 types [4].

1. Correlation Measurement Under the assumption that a good feature subset contains high feature-class relevance and low inter-feature relevance, some previous research works focused on measuring relevance using a linear correlation coefficient. The Pearson correlation coefficient is one of those simple and popular measurements, as shown below:

$$R(i) = \frac{cov(X_i, Y)}{\sqrt{(var(X_i)var(Y)}}$$

2. Distance Measurement Like unsupervised learning algorithms, distance-based feature selection is based on the assumption that good feature subset can make the distance between samples belonging to the same class as small as possible, and samples that

belong to different classes as far as possible. Some common distance measurements include Euclidean distance, Mahalanobis distance, Minkowski distance etc.

3. Information Measurement Measurements such as mutual information [14],[15] and information gain are based on the simple assumption that the more information we gain, the better a feature subset is. Information gain is a concept of information theory, which originates from the concept of entropy (also a measurement of uncertainty):

$$H(Y) = \sum_{i=1}^{m} P_i \log_2 P_i$$

After adding condition variable X ($X = x_i$), the conditional entropy of Y is:

$$H(Y) = \sum_{i=1}^{m} P_{x=x_i} H(Y|X = x_i)$$

After adding conditional X, the uncertainty of Y has decreased. The decrease in uncertainty is an increase in certainty, called information gain, calculated as follows:

$$IG(Y|X) = H(Y) - H(Y|X)$$

4. Consistency Measurement Consistency-based feature selection methods follow the simple assumption that good feature subsets should consist of as few features as possible while maintaining high consistency at the same time. i.e. if two samples contain the same values of a certain feature set, but belong to different classes, then this feature set should not be considered as an optimal subset. Regarding the interference of noise, it's more reasonable to set an inconsistency rate to avoid ignoring good features. One example is rough set theory [14, 16, 17, 18].

Wrapper Method:

For wrapper methods, feature selection is integrated into training process [19]. Acting as classifiers, wrappers classify feature sets using the selected subset. Measurements of model

forecasting capability, such as classification accuracy, serve as evaluation criteria of the quality of selected subsets. The evaluation process of wrapper methods depends on specific inductive algorithm, which in return limits its portability to other algorithms. Although the quality of feature subsets generated by wrapper methods generally outperforms that of filter methods, the much heavier computational cost should also be carefully bargained.

Hybrid Method:
With the objective of reducing the computational cost of wrapper methods while maintaining their outstanding classification accuracy, several research works were carried out to seek the possibility to speed up the convergence of wrapper methods by combining filter methods. Several hybrid methods were proposed in [13, 20, 21, 22] to absorb the advantages of both types.

Label-based category
Based on whether data are labeled, feature selection methods can also be generally divided into supervised feature selection and unsupervised feature selection. Since the quality of the original data set has great impacts on future performance of machine learning algorithms, the lack of class labels brings more difficulties for feature selection since less natural grouping information can be acquired from the original data. Especially in nowadays' big data environment, Internet traffic researchers face bigger challenges, because Internet traffic data tend to be more structurally complex, vague, label-missing and of greater scale. Several supervised feature selection methods are listed [23, 24, 25, 26, 27], and a series of valuable works on unsupervised feature selection methods can be found in [28, 29, 30, 31, 32, 33, 34].

5.2.3 Feature Selection Methods in Internet Traffic Characterization

In the research domain of Internet traffic characterization, feature selection keeps attracting special attention because the scale of Internet traffic feature sets has experimented a rapid explosion during the last decade. Work in [35] has shown the effectiveness of feature se-

lection to improve computational performance without severely reducing classification accuracy when conducting traffic flow identification. They undertook two different evaluation measurements- consistency and correlation, and further compared greedy and best first (both forward and backward directions) search strategies. [1] generated an original traffic set that contains 248 features, on which direct importation to classification models is infeasible due to the existence of high irrelevance and redundancy, and a great computational cost. Fast Correlation-Based Filter, proposed in [36], together with a novel wrapper-based method to determine threshold, were used to select useful features. The same data sets were reused in [37], which proposed a new feature selection method called BFS. BFS is more competitive in maintaining the balance of multi-class classification results in comparison with FCBF regarding metrics like g-mean and classification accuracy. [38] proposed an application-based feature subset selection using parameter estimation for each logistic regression model established for the corresponding application class. This can effectively resolve the imbalance problem caused by elephant flows. [39] discussed real-time Internet traffic identification that has special requirements of simplicity and effectiveness on feature subsets. [40] performed Correlation based, Consistency based, and PCA [41] feature selection on a real-time Internet traffic dataset obtained by a packet capture tool. [42] proposed a mutual-information-based feature selection and automatic determiner of the number of relevant features. An outlier detection method which improved PCA was also brought out to avoid the influence of outliers in real traffic. In [43] new evaluation metrics named goodness, stability and similarity were used to assess the advantages and defects of existing feature selection methods. Then they integrated six relatively competitive FS methods (Information Gain, Gain Ratio, PCA, Correlation-Based Feature Selection, Chi-square, and Consistency-Based Feature Selection) to combine their strengths. This new integrated technique consists of two procedures: first, a consistent subset generated from the results of different feature selection methods was discovered; then based on this candidate subset, a proposed measurement of support was used to obtain the final feature subset. [44] first proposed a novel feature selection metric named Weighted Symmetrical Uncertainty (WSU), then brought about a hybrid FS method that prefilters features using WSU. Later it adopted a wrapper method that selected features regarding the Area Under roc Curve (AUC) metric. An additional algorithm SRSF was also proposed to

get rid of the effect of dynamic traffic flows. [45] proposed a novel feature selection method that made use of both linear correlation coefficient measurement and non-linear mutual information measurement. Zulaiha et. Al first proposed a wrapper method in [46] using Bees Algorithm as search strategy and Support Vector Machine as the classifier. After comparison, it turned out that BA yielded better results than other FS methods such as Rough-DPSO, Rough Set, Linear Genetic Programming, MARS and Support Vector Decision Function Ranking. Then in [47] they proposed a hybrid FS method called LGP_BA which is a combination of Linear Genetic Programming and Bee Algorithm that achieved better accuracy and efficiency.

5.3 Future Practical Application and Conclusions

The ONTIC (Online Network TraffIc Characterization) project, funded by the seventh Framework Programme for Research and Technological Development of European Commission, aims for accurate identification and characterization of network traffic regarding different application types, which will significantly contribute to problems such as dynamic QoS management, network intrusion detection and fast detection of network congestion [48]. To achieve this objective, a Peta Byte size data set composed of real network traffic summaries will be generated from several data flows. These flows will be captured in the core network of an ISP for several months. Making use of such an adequate raw data set, more than two hundred features are generated to form the feature space with the help of a TCP sTatistic and Analysis Tool - Tstat [49]. These features are used to characterize Internet traffic flows and inherently classify them into corresponding Internet traffic classes, which are often of different granularities: it can be simply divided into normal and abnormal classes, or various specific application categories in terms of detailed circumstances. There will be more than 240 features in our raw feature set; therefore, it's necessary to conduct a feature selection procedure to reduce to expected scale (in our case 5-8 features are sufficient). Since different feature selection methods adopt specific and independent criteria to perform the feature pruning step, generated subsets are probably discriminative and of different sizes, which makes direct comparison meaningless. Several feature selection methods will be selected and implemented towards the original full feature set, and a thorough comparison of experiment

results regarding some generic metrics like classification accuracy and running speed will be conducted making use of different methods. In this paper we provide an introduction to feature selection frameworks and different categories of methods. Several algorithms are briefly explained, including both classic and newly-proposed methods. Then feature selection in the domain of Internet traffic characterization is especially discussed. Considering the variety of existing methods and rapid proposal of new methods, this survey is far from complete. Future application of feature selection to ONTIC project is also proposed to verify the practical significance of feature selection.

Bibliography

[1] Moore, A., Zuev, D., “Internet Traffic Classification Using Bayesian Analysis Techniques”, In ACM SIGMETRICS Performance Evaluation Review, vol. 33, no. 1, pp. 50–60, 2005.

[2] Choudhary, A., Kumar, J., “Survey on Hybrid Approach for Feature Selection”, International Journal of Science and Research, vol. 3, issue 4, pp. 438–439, 2014.

[3] Dash, M., Liu, H., “Feature Selection for Classification”, Intelligent Data Analysis, vol. 1, issues 1–4, pp. 131–156, 1997.

[4] Liu, H., Yu, L., “Toward Integrating Feature Selection Algorithms for Classification and Clustering”, IEEE Transactions on Knowledge and Data Engineering, vol. 17, issue 4, pp. 491–502, 2005.

[5] Song, Q., Ni, J., Wang, G., “A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, issue 1, pp. 1–14, 2013.

[6] Molina, L., Belanche, L., Nebot, A., “Feature Selection Algorithms: A Survey and Experimental Evaluation”, Proceedings International Conference on Data Mining (ICDM 2002), pp. 306–313, 2002.

[7] Yu, L., Liu, H., “Efficient Feature Selection via Analysis of Relevance and Redundancy”, Journal of Machine Learning Research, vol. 5, pp. 1205–1224, 2004.

[8] Guyon, I., Elisseeff, A., “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[9] Chandrashekar, G., Sahin, F., "A survey on feature selection methods", Computers and Electrical Engineering, vol. 40, issue 1, pp. 16–28, 2014.

[10] Steinbiss, V., Tran, B., Ney, H., "Improvements in Beam Search", Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP1994), vol. 94, no. 4, pp. 2143–2146, 1994.

[11] Cantu-Paz, E., "Feature Subset Selection by Estimation of Distribution Algorithms", Proceedings of the Genetic and Evolutionary Computation Conference (GECCO2002), vol. 2, pp. 303–310, 2002.

[12] Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., Wang, S., "An Improved Particle Swarm Optimization for Feature Selection", Journal of Bionic Engineering, vol. 8, issue 2, pp. 191–200, 2011.

[13] Cantu-Paz, E., Newsam, S., Kamath, C., "Feature Selection in Scientific Applications", Proceedings of the 10th ACM International conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 788–793, 2004.

[14] Foithong, S., Pinngern, O., Attachoo, B., "Feature subset selection wrapper based on mutual information and rough sets", Expert Systems with Applications, vol. 39, Issue 1, pp. 574–584, 2012.

[15] Peng, H., Long, F., Ding, C., "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, issue 8, pp. 1226–1238, 2005.

[16] Bae, C., Yeh, W., Chung, Y., Liu, S., "Feature selection with Intelligent Dynamic Swarm and Rough Set", Expert Systems with Applications, vol. 37, issue 10, ". 7026–7032, 2010.

[17] Qian, Y., Liang, J., Pedrycz, W., Dang, C., "Positive approximation: An accelerator for attribute reduction in rough set theory", Artificial Intelligence, vol. 174, issues 9–10, pp. 597–618, 2010.

[18] Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R., "Feature selection based on rough sets and particle swarm optimization", Pattern Recognition Letters, vol. 28, issue 4, pp. 459–471, 2007.

[19] Kohavi, R., John, G., "Wrappers for feature subset selection", Artificial Intelligence, vol. 97, issues 1-2, pp. 273–327, 1997.

[20] Bermejo, P., Ossa, L., Gámez, J., Puerta, J., "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", Knowledge-Based Systems, vol. 25, issue 1, pp. 35-44, 2012.

[21] Naidu, K., Dhenge, A., Wankhade, K., "Feature Selection Algorithm for Improving the Performance of Classification: A Survey", Fourth International Conference on Communication Systems and Network Technologies, Communication Systems and Network Technologies (CSNT), pp. 468–471, 2014.

[22] Gheyas, I., Smith, L., "Feature subset selection in large dimensionality domains", Pattern Recognition, vol. 43, issue 1, pp. 5–13, 2010.

[23] Guan, Y., Dy, J., Jordan, M., "A Unified Probabilistic Model for Global and Local Unsupervised Feature Selection", Proceedings of the 28th International Conference on Machine Learning (ICML 11), pp. 1073–1080, 2011.

[24] Tang, J., Liu, H., "Unsupervised Feature Selection for Linked Social Media Data", Proceedings of the 18th ACM International conference on Knowledge discovery and data mining (KDD 12), pp. 904–912, 2012.

[25] Cai, D., Zhang, C., He, X., "Unsupervised Feature Selection for Multi-Cluster Data", Proceedings of the 16th ACM International conference on Knowledge discovery and data mining (KDD 2010), pp. 333–342, 2010.

[26] Qian, M., Zhai, C., "Unsupervised Feature Selection for Multi-View Clustering on Text-Image Web News Data", CIKM2014, 2014.

[27] Azizyan, M., Singh, A., Wasserman, L., "Feature Selection for High-Dimensional Clustering", Cornell University Library, 2014.

[28] Aliakbarian, M.S. , Fanian, A. , Saleh, F.S. , Gulliver, T.A., "Optimal supervised Feature extraction in Internet traffic classification", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2013.

[29] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., Kannan, A., "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey", EURASIP Journal on Wireless Communications and Networking, 2013.

[30] Tsang, I., Tan , M., Wang, L., "Learning Sparse SVM for Feature Selection on Very High Dimensional Datasets", Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 1047–1054, 2010.

[31] Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X., "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation", Journal of Machine Learning Research, January , vol. 11, pp. 171–234, 2010.

[32] Sotoca, J., Pla, F., "Supervised feature selection by clustering using conditional mutual information-based distances", Pattern Recognition, vol. 43, issue 6, pp. 2068–2081, 2010.

[33] Song, L., Smola, A., Gretton, A., Borgwardt, A., Bedo, J., "Supervised Feature Selection via Dependence Estimation", 24th International Conference on Machine Learning (ICML 2007), 2007.

[34] Jing-jing, Z., Xiao-hong, H., Qiong, S., Yan, M., "Real-time feature selection in traffic classification", Journal of China Universities of Posts and Telecommunications, vol. 15, pp. 68–72, 2008.

[35] Williams, N., Zander, S., Armitrage, G. "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification", ACM SIGCOMM Computer Communication Review, 2006.

[36] Yu, L., Liu, H., "Feature selection for high-dimensional data: a fast correlation-based filter solution", Conf. Machine Learning (ICML 03), 2003.

[37] Zhen, L., Qiong, L., "A New Feature Selection Method for Internet Traffic Classification Using ML". Physics Procedia Internet Conference on Medical Physics and Biomedical Engineering (ICMPBE 2012), 2012.

[38] En-Najjary, T., Urvoy-Keller, G., Pietrzyk, M., "Application-based Feature Selection for Internet Traffic Classification", 22nd International Teletraffic Congress, 2010.

[39] Chen, Z., Peng, L., Zhao, S., Zhang, L., Jing, S., "Feature Selection Toward Optimizing Internet Traffic Behavior Identification, Algorithms and Architectures for Parallel Processing", Lecture Notes in Computer Science, vol. 8631, pp. 631–644, 2014.

[40] Kuldeep, S., Agrawal, S., "Performance Evaluation of Five Machine Learning Algorithms and Three Feature Selection Algorithms for IP Traffic Classification", IJCA Special Issue on Evolution in Networks and Computer Communications, vol. 1, pp. 25–32, 2011.

[41] Ding, C., Zhou, C., He, X., Zha, H., "R1-PCA: Rotational Invariant L1-norm Principal Component Analysis for Robust Subspace Factorization", 23rd International Conference on Machine Learning (ICM L2006), 2006.

[42] Pascoal, C., Rosario de Oliveira, M., Valadas, R, Filzmoser, P., Salvador, P., Pacheco, A. "Robust Feature Selection and Robust PCA for Internet Traffic Anomaly Detection", Proceedings of IEEE INFOCOM, 2012.

[43] Fahad, A., Tari, Z., Khalil, I., Habib, I., Alnuweiri, H., "Toward an efficient and scalable feature selection approach for Internet traffic classification", Computer Networks, vol. 57, isuue 9, pp. 2040–2057, 2013.

[44] Zhang, H., Lu, G., Qassrawi, M., Zhang, Y., Yu, X., "Feature selection for optimizing traffic classification", Computer Communications, vol. 35, issue 12, pp. 1457-1471, 2012.

[45] Amiri, F., Yousefi, M., Lucas, C., Shakery, A., Yazdani, N., "Mutual information-based feature selection for intrusion detection systems", Journal of Network and Computer Applications, vol. 34, issue 4, pp. 1184–1199, 2011.

[46] Alomari, O., Othman, Z., "Bees Algorithm for feature selection in Network Anomaly detection", Journal of Applied Sciences Research, vol. 8, issue 3, pp. 1748–1756, 2012.

[47] Hasani, S.R., Z.A. Othman and S.M.M. Kahaki, "Hybrid feature selection algorithm for intrusion detection system", Journal of Computer Science, vol. 10, issue 6, pp. 1015–1025, 2014.

[48] ONTIC, Online Network Traffic Characterization, http://www.ict-ontic.eu/

[49] Tstat, TCP STatistic and Analysis Tool, http://tstat.tlc.polito.it/index.shtml

BIBLIOGRAPHY

BIBLIOGRAPHY